The Devil Is in the Frequency: Geminated Gestalt Autoencoder for Self-Supervised Visual Pre-Training

Hao Liu[†*], Xinghua Jiang[*], Xin Li, Antai Guo, Yiqing Hu, Deqiang Jiang, Bo Ren

Tencent YouTu Lab

{ivanhliu, clarkjiang, fujikoli, ankerguo, hooverhu, dqiangjiang, timren}@tencent.com

Abstract

The self-supervised Masked Image Modeling (MIM) schema, following "mask-and-reconstruct" pipeline of recovering contents from masked image, has recently captured the increasing interest in the community, owing to the excellent ability of learning visual representation from unlabeled data. Aiming at learning representations with high semantics abstracted, a group of works attempts to reconstruct non-semantic pixels with large-ratio masking strategy, which may suffer from "over-smoothing" problem, while others directly infuse semantics into targets in off-line way requiring extra data. Different from them, we shift the perspective to the Fourier domain which naturally has global perspective and present a new Masked Image Modeling (MIM), termed Geminated Gestalt Autoencoder (Ge^2-AE) for visual pre-training. Specifically, we equip our model with geminated decoders in charge of reconstructing image contents from both pixel and frequency space, where each other serves as not only the complementation but also the reciprocal constraints. Through this way, more robust representations can be learned in the pre-trained encoders, of which the effectiveness is confirmed by the juxtaposing experimental results on downstream recognition tasks. We also conduct several quantitative and qualitative experiments to investigate the learning behavior of our method. To our best knowledge, this is the first MIM work to solve the visual pre-training through the lens of frequency domain.

Introduction

Recently, self-supervised visual pre-training has witnessed a fast-paced progress in learning robust representations of visual content, which can overcome the *data appetite* problem suffered by supervised learning paradigm, where large amount of labeled training data is required. Among, a series of methods, termed Masked Image Modeling (MIM) (He et al. 2021; Wei et al. 2021; Xie et al. 2021; Dong et al. 2021; Zhou et al. 2021b; Chen et al. 2022), exhibits promising potential, which inherits the "mask-and-reconstruct" thought from masked autoencoding methods in natural language processing (NLP) field, such as BERT(Devlin et al. 2018). More concretely, parts of content in input image are masked to learn latent representations from the visible regions by encoder, which are then used to reconstruct content by decoder. For the downstream visual tasks, the encoder is expected to yield robust representation with *rich semantic information*. However, the language naturally has semantics highly abstracted by human (He et al. 2021), where the elemental unit is also discrete. In contrast, pixels contained in image merely contain low-level statistics and often present the heavy redundancy due to the continuous property of image. Therefore, the targets setting become the crucial component in the MIM, which essentially determine what unwanted minutiae should be compressed for the semantic-perceiving purpose.

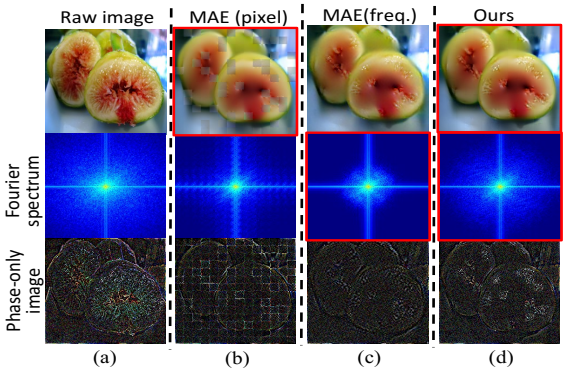


Figure 1: Demonstrations of reconstructed images (1st. row), Fourier spectrum maps (2nd. row) and phase-only images (3rd. row) yielded by MAE and our proposed method. The (a) column is the raw image and corresponding maps, while (b) and (c) are results of MAE regarding pixel and frequency as targets. (d) are ours. The direct predictions are highlighted by red boundaries, and their corresponding pixel or frequency maps are obtained by 2D-FFT or 2D-IFFT. The phase-only images indicating semantics are obtained by setting the amplitude component to a constant.

To implement the purpose, existing MIM methods have made attempts mainly in two directions *w.r.t* the target setting. Among, MAE (He et al. 2021) and SimMIM (Xie et al. 2021) are two simple and straightforward methods, which regard raw pixels as targets and encourage the model concentrating on semantics by large-ratio masking strategy. Although ViT (Dosovitskiy et al. 2020) is adopted, the target pixels serving as non-semantic entities still contain heavy

[*]Equal contribution. [†]Contact person.

redundancy and lack global perspective, which is essential to the high semantics capture. Alternatively, extra knowledge is introduced in targets to achieve the goal in tokenizer-based methods. Nevertheless, there is no free lunch. These methods reconstruct content supervised by off-line or on-line trained tokenizers (Bao, Dong, and Wei 2021; Zhou et al. 2021b; Dong et al. 2021), which may inevitably incur complicated processing, such as knowledge distillation or contrastive learning (Zhou et al. 2021b). Although MaskFeat (Wei et al. 2021) simply regards handcrafted HOG feature as target (Wei et al. 2021), the target is still a local statistic-based descriptor, which may also suffer from the same problem with pixel prediction-based methods. Conclusively, a qualified supervision should present the following properties: 1) *well discretizing continuous image content*; 2) *containing high visual semantic*. We ask: "*Is there a simple way to obtain the supervision with the desired properties merely from image per se?*"

To answer this question, we shift our perspective to the frequency domain. Our motivation comes from several intriguing properties of the Fourier transformation. Firstly, as suggested by many works (Oppenheim et al. 1979; Oppenheim and Lim 1981; Piotrowski and Campbell 1982; Hansen and Hess 2007), high-level semantics of the original signal are naturally preserved in the phase component of Fourier spectrum, while the amplitude component contains low-level statistics. It can be vividly demonstrated by the clear object contour in the phase-only image (last image of Fig. 1(a)), where the amplitude component is set to a constant. Besides, Fourier basis are orthogonal, which can spontaneously discretize spatial domain content into individual frequency component in Fourier spectrum. Last but not least, each frequency component in Fourier spectrum has the global vision, which of each is also, to put it another way, a high summary of whole image.

By visualizing the Fourier spectrum of recovered images of MAE (Fig. 1(b)), we find that a portion of high frequency is compressed while low frequency is dominant. Although the phase-only image exhibits semantic contour to certain extent, there are also some unwanted details preserved, *e.g., grid like texture caused by independent pixel reconstruction process*. As a result, the integrity of object contour would be disrupted. Alternatively, we come up with a solution by directly replacing the pixel target in MAE (He et al. 2021) with the Fourier spectrum where each component carries the global information. From the phase-only image of Fig. 1(c), although a more holistic contour can be observed, the pre-trained encoder still can not yield decent representation on downstream tasks. We attribute it to the "over-smoothing" issue, *e.g.*, too many useful information hidden in high or middle frequencies are abandoned, as illustrated by Fourier spectrum in Fig. 1(c). That is to say, the model overly pays attention to semantics while certain local details are ignored, whereas a good representation needs both semantic and local details, as suggested by work (Dong et al. 2021). To achieve this goal and substantially retain the merit of "simple yet effective", we build our MIM method, termed **Ge**minated **Ge**stalt **A**uto**E**ncoder (Ge^2-AE), upon canonical MAE (He et al. 2021) and simply modify it with one extra lightweight frequency decoder (FD) added to simultaneously perform gestalt tasks of the local masked region and global frequency. To mitigate the inconsistency between pixel and frequency space, the task-specific FD is equipped with Fourier Spectrum Perceiver (FSP) and process spatial-frequency contents alternatively to adapt the spatial encoded tokens to the frequency prediction task. Moreover, the information in pixel and frequency space also play as the reciprocal constraints dressed on each other when reconstructing contents from both domains. As illustrated by Fig. 1(d), despite that the reconstructed image of Ge^2-AE has no obvious difference from MAE, the yielded Fourier spectrum overcomes the "over-smoothing" problem with more proper details and holistic object contour (refer to phase-only image in Fig. 1(d)) preserved. Benefiting from the tailored design, our Ge^2-AE can achieve significant performance improvement on pre-trained representation quality than other methods, and shows inspiring transferability, which are validated by the experimental results on several downstream visual recognition tasks. In summary, our contributions are:

- We reinspect the Masked Image Modeling for visual pre-training task from frequency domain and investigate intriguing properties for robust representation learning. To our best knowledge, we are the first to explore the space domain imperceptible clues from frequency perspective in visual pre-training field.
- We coin a novel Ge^2-AE tailored for visual pre-training problem, consisting of geminated decoders performing gestalt from both spatial and frequency domain, which can fully leverage their reciprocity to enhance the semantics without introducing extra complicated processes.
- Quantitative and qualitative experimental results on public benchmarks demonstrate that our method achieve significant performance improvement on pre-trained representation quality than state-of-the-arts.

The Methodology

Preliminary: 2D-Discrete Fourier Transform

Before introducing our method, we first briefly review the 2D-Discrete Fourier Transform (DFT) serving as an important tool in signal analysis, which plays an indispensable role in our Ge^2-AE. Given a 2D signal (one channel of image or feature cube) $\mathbf{F} \in \mathbb{R}^{W \times H}$, its 2D-DFT can be defined as:

$$f(u,v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} F(h,w) e^{-j2\pi\left(\frac{uh}{H} + \frac{vw}{W}\right)}, \qquad (1)$$

where $F(h,w)$ represents the h-th and w-th pixel or element in $\mathbf{F}$ while the u and v are indexes of horizontal and vertical spatial frequencies in Fourier spectrum. Correspondingly, the 2D Inverse DFT (2D-IDFT) is formulated as:

$$F(h,w) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} f(u,v) e^{j2\pi\left(\frac{uh}{H} + \frac{vw}{W}\right)}, \qquad (2)$$

Both DFT and IDFT can be accelerated with the fast version, FFT algorithm (Nussbaumer 1981). Moreover, the definitions of amplitude A and phase components P deduced

from Fourier spectrum are given as:

$$A(u,v) = \left[\mathcal{R}^2(u,v) + \mathcal{I}^2(u,v)\right]^{1/2}, \tag{3}$$

$$P(u,v) = \arctan\left[\frac{\mathcal{I}(u,v)}{\mathcal{R}(u,v)}\right], \tag{4}$$

where $\mathcal{R}$ and $\mathcal{I}$ denote the real and imaginary part of frequency f respectively, *i.e.*, $f(u,v) = \mathcal{R}(u,v) + \mathcal{I}(u,v)i$.

Architecture of Ge^2-AE

The architecture of our proposed Ge^2-AE is illustrated in Fig. 2(a), which is built upon a representative baseline, Masked Autoencoder (MAE). Specifically, we inherit the encoder design of MAE, where the canonical ViT is adopted to project the unmasked patches into visible tokens. As suggested by MAE (He et al. 2021), the decoder design is crucial to the MIM model, as it not only learns representations of masked tokens, but also determines the semantic level of the whole learned latent representations.

The pre-training task-tailored designs of our method lie in the following three folds: **I) Geminated structure**. Considering the raw pixels regarded as targets in MAE are non-semantic entities, we simply replace pixel targets with Fourier spectrum containing global semantics. However, we observe the evident performance drops. Therefore, we equip our MIM pre-training framework with decoupled decoders in charge of reconstructing spatial and frequency domain simultaneously. The assumption behind our design is that *a good latent representation should be a consensus between local details and global semantics.* We preserve the decoder from the vanilla MAE as the pixel decoder (PD) for the similar purpose, which deploys ViT blocks as the core ingredients. **II) Frequency decoder**. We modify the vanilla ViT blocks to adapt to the frequency gestalt task, which will be elaborated in "Frequency Decoder" Section. **III) Complementary constraints.** The information of spatial and frequency domain are not only expected to be predicted, but also can provide the complementary constraints for each other. As shown in Fig. 2(a), taking the pixel reconstruction branch for example, we further convert the predicted image to Fourier domain by FFT and treat the Fourier spectrum of raw image as constraint for it. To put it in another way, this constraint encourages the global frequency involving semantics preserved at most when reconstructing image. Correspondingly, the pixel constraint put on reconstructed Fourier spectrum aims to preserve more local content details in the predicted spectrum.

Frequency Decoder

As described above, the frequency decoder (FD) in our method plays a role in infusing the global frequency information to the learned latent representation. A plain way to implement it is to apply FFT either on the input or the output of decoder. However, neither the early-FFT nor late-FFT process can well adapt to the task of directly predicting global frequency from spatial encoded tokens, not to mention that some of them are masked. Alternatively, we consider to achieve the above purpose in a progressive way to alleviate the inconsistency problem of spatial-frequency domain. Inspired by the frequency usage in NomMer (Liu et al. 2021), as demonstrated in Fig. 2(b), we make several task-specific modifications on the vanilla ViT blocks. Specifically, the basic block takes spatial features as input appended by a Layer Normalization (LN) (Vaswani et al. 2017), then it is processed by the 2D-FFT (purple arrow) defined in Eqn. (1). Once converted to the frequency spectrum $\tilde{\mathbf{f}} \in \mathbb{R}^{W\times H\times C}$, each component in it can have global vision on the whole image but comprised by different vision patterns, *e.g.,* texture of dog hairs. To highlight those significant frequency components, which could be propagated to the learned latent representations and informative for visual recognition, we insert a Fourier Spectrum Perceiver (FSP) inside FD, formulated as $\mathbf{f} = \mathbf{\Omega} \odot \tilde{\mathbf{f}}$,

where $\mathbf{\Omega} \in \mathbb{R}^{W\times H\times C}$ is the learnable parameter matrix. "$\odot$" represents the Hadamard product. Once the emphasized frequency $\mathbf{f}$ obtained, the 2D-IFFT (orange arrow) defined in Eqn. (2) is applied to restore the spatial feature $\mathbf{F}$. Afterwards, Feed-Forward Networks (FFN) (Vaswani et al. 2017) equipped with residual connection and LN are appended, which is similar to that of vanilla ViT block. These basic blocks can be stacked L times, enabling the alternative reconstruction of spatial-frequency contents, which can feasibly solve the inconsistency problem. In the end, the 2D-FFT with LN is applied to predict the final reconstructed Fourier spectrum.

Moreover, it is worth mentioning that, compared to vanilla ViT blocks deploying self-attention operation with $\mathcal{O}(n^2)$ complexity (n is the token number), our FD is a lightweight module. In detail, the total computational complexity of FD's basic block is $\mathcal{O}(n \log n)$, where 2D-FFT and 2D-IFFT are both with $\mathcal{O}(n \log n)$ complexity while that of Hadamard product is only $\mathcal{O}(n)$.

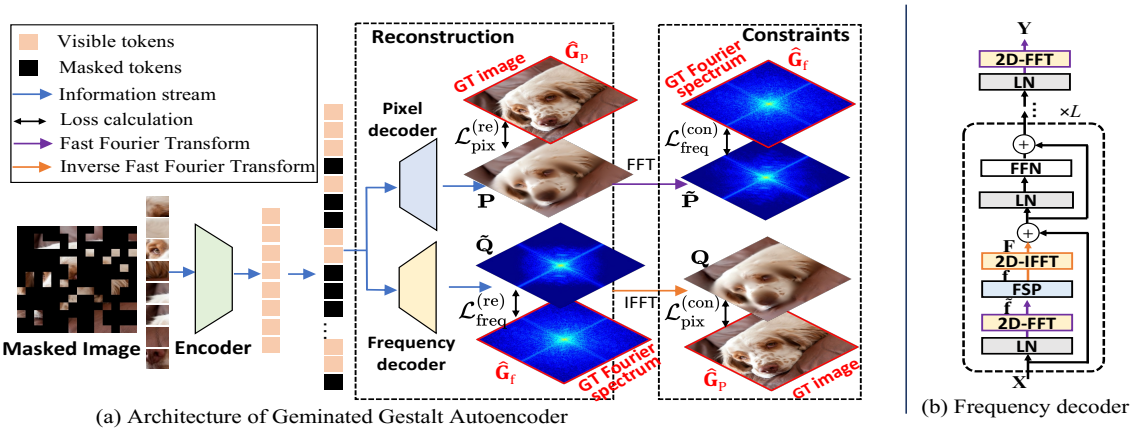
Pre-training Strategy

Overall Loss. During pre-training, our Ge^2-AE learns latent representation by solving content gestalt from both pixel-level and frequency-level:

$$\begin{aligned}\mathcal{L} = &\overbrace{\mathcal{L}^{(\text{re})}_{\text{pix}}(\mathbf{P}, \hat{\mathbf{G}}_{\text{p}}) + \mathcal{L}^{(\text{con})}_{\text{freq}}(\tilde{\mathbf{P}}, \hat{\mathbf{G}}_{\text{f}})}^{pixel-level} \\ &+ \underbrace{\lambda(\mathcal{L}^{(\text{re})}_{\text{freq}}(\tilde{\mathbf{Q}}, \hat{\mathbf{G}}_{\text{f}}) + \mathcal{L}^{(\text{con})}_{\text{pix}}(\mathbf{Q}, \hat{\mathbf{G}}_{\text{p}}))}_{frequency-level},\end{aligned} \tag{5}$$

where $\mathcal{L}^{(re)}_{\sim}$ represent losses for reconstruction task while $\mathcal{L}^{(con)}_{\sim}$ serve as constraints, as illustrated in Fig. 2(a). $\mathbf{P}$ and $\tilde{\mathbf{Q}}$ are the predictions of pixel and frequency decoders while $\tilde{\mathbf{P}}$ and $\mathbf{Q}$ are their corresponding transformations obtained by 2D-FFT and 2D-IFFT. $\hat{\mathbf{G}}_{\text{p}}$ and $\hat{\mathbf{G}}_{\text{f}}$ are the raw image and its Fourier spectrum serving as Ground Truths (GTs) for both reconstruction and subsequent constraints. The effect of either decoder is controlled by loss weight λ (λ =0.5 in default). For the pixel loss $\mathcal{L}^{\sim}_{\text{pix}}$, we compute Mean Square Error (MSE) between the reconstructed and raw images in pixel space, which is similar to MAE (He et al. 2021).

Frequency loss. As studied in previous works (Rahaman et al. 2019; Xu, Zhang, and Luo 2022; Wang et al.



(a) Architecture of Geminated Gestalt Autoencoder (b) Frequency decoder

Figure 2: Architecture of the Proposed Geminated Gestalt Autoencoder (Ge^2-AE). The encoder receives the unmasked patches to yield visible tokens, which are sent to the geminated structure decoders together with masked tokens to recover in pixel and Fourier domain constrained by each other. Best viewed in color and zoom in.

2020) on the learning behavior from frequency domain, *spectral bias* of the deep neural networks is often inclined to low frequency functions. Besides, according to *F-Principle* (Xu et al. 2019), the fitting priority of a network to certain frequencies is various throughout the training, often in the low-to-high pattern. As a result, for our frequency decoder, some significant frequency could be hardly decoded, once those frequencies with higher priority are generated. Specifically, high-priority frequency is termed *easy frequency*, otherwise *hard frequency*.

To better capture the hard frequency, we adopt focal frequency loss (Jiang et al. 2021) as $\mathcal{L}_{\text{freq}}$ to dynamically tune each frequency weight, which can be defined as:

$$\mathcal{L}_{\text{freq}} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\omega(u,v)\odot\gamma(f(u,v),\hat{f}(u,v))^2, \tag{6}$$

where $f(u,v)$ is the i,j-th frequency component of spectrum $\tilde{\mathbf{Q}}$ or $\tilde{\mathbf{P}}$ while $\hat{f}(u,v)$ denotes the frequency component from GT spectrum $\hat{\mathbf{G}}_{\text{freq}}$ at the same location. In addition, $\gamma(f,\hat{f})$ is the frequency distance implemented by computing squared Euclidean distance between their real and imaginary parts. ω is spectrum weight matrix can down-weight the easy frequencies, which are formulated as:

$$\omega(u,v) = \gamma(f(u,v),\hat{f}(u,v))^{\beta}, \tag{7}$$

$$\gamma(f,\hat{f}) = \sqrt{(\mathcal{R}-\tilde{\mathcal{R}})^2+(\mathcal{I}-\tilde{\mathcal{I}})^2}, \tag{8}$$

β is the scaling factor for flexibility, which is set to 1 in default.

Experiments

Implementation Details

For the pre-training on ImageNet-1K (IN1K) training set, we inherit the experimental settings in MAE (He et al. 2021). The random mask ratio is set to 75% in default while the block number L is empirically set to 8 for both PD and FD. Moreover, with the initial learning rate and batch size are set to $1.5e^{-4}$ and 4,096, respectively, all our models are pre-trained for 800 epochs with the input size of 224^2. Besides, AdamW (Loshchilov and Hutter 2017) optimizer with a cosine learning rate scheduler is adopted. We implement the Ge^2-AE architecture by Pytorch (Paszke et al. 2019) with all experiments conducted on a workstation with 32 NVIDIA A100-40 GB GPUs. All the reported results are the averaged ones over 10 random seeds.

Method	ViT-S	ViT-B	ViT-L	ViT-H
DINO (Caron et al. 2021)	-	82.8	-	-
CAE (Chen et al. 2022)	81.8	83.6	-	-
CIM (Fang et al. 2022)	81.6	83.1	-	-
BEiT (Bao, Dong, and Wei 2021)	-	83.2	85.2	-
MoCo v3 (Chen, Xie, and He 2021)	-	83.2	84.1	-
MaskFeat (Wei et al. 2021)	-	84.0	85.7	-
iBOT (Zhou et al. 2021b)	**82.3**	84.0	85.2	-
PeCo (Dong et al. 2021)	-	84.5	86.5	87.5
SimMIM (Xie et al. 2021)	-	83.8	-	-
MAE (He et al. 2021)	-	83.6	85.9	86.9
Ge^2-AE (Ours)	82.2	**84.8**	**86.6**	**87.7**

Table 1: ImageNet-1K (IN1K) fine-tuning Top-1 accuracy of ViTs with different sizes.

Image Classification on ImageNet-1K

We firstly perform self-supervised pre-training on the ImageNet-1K (IN1K) (Deng et al. 2009) training set and evaluate the learned representation quality with only the encoder preserved under two supervised training settings: **1) end-to-end fine-tuning; 2) linear probing**.

Experimental setting. For the end-to-end fine-tuning, our model employing ViT-base as default encoder is trained for 300 epochs with the input size of 224^2, of which the initial learning rate and batch size are set to $1e^{-3}$ and 1,024, respectively. Besides, AdamW (Loshchilov and Hutter 2017) optimizer with a cosine learning rate scheduler is adopted.

Method	Arch.	Param.	Epoch	Linear
DINO (Caron et al. 2021)	ViT-S	21	-	77.0
DINO (Caron et al. 2021)	ViT-B	85	-	78.2
MoCo v3 (Chen, Xie, and He 2021)	ViT-B	86	-	76.7
CAE (Chen et al. 2022)	ViT-S	21	300	50.8
CAE (Chen et al. 2022)	ViT-B	85	800	68.3
iBOT (Zhou et al. 2021b)	ViT-S	21	3,200	77.9
iBOT (Zhou et al. 2021b)	ViT-B	85	1,600	**79.5**
SimMIM (Xie et al. 2021)	ViT-B	85	-	56.7
MAE (He et al. 2021)	ViT-B	86	800	68
Ge^2-AE (Ours)	ViT-B	86	800	75.3

Table 2: ImageNet-1K (IN1K) Top-1 accuracy of different methods under linear probing setting.

The weight decay is set to 0.05 and the maximal gradient norm is clipped to 5.0. As for the linear probing setting, we exploit the LARS (You, Gitman, and Ginsburg 2017) optimizer with batch size 16,384 and 0.1 learning rate, and train the model for 100 epochs.

End-to-end fine-tuning. Under this setting, the pre-trained encoder is fine-tuned with the classification head together. From the Tab. 1, one can observe that our Ge^2-AE can achieve round 1% performance improvement than the baseline method MAE. Compared to those MIM methods (PeCo, iBOT) introducing complex contrastive operations yield tokenizers, our method can also achieve the competitive results for ViTs in different model sizes, but with only a lightweight frequency decoder added. We attribute it to the good global semantic perceiving ability of our method, which is comparable to the tokenizer-based methods.

Linear probing. When evaluated under linear probing setting, all parameters of the pre-trained encoder are frozen while only the last classification layer is trained. From Tab. 2, we surprisingly find that our method can surpass the MAE by round 7% with the similar pre-training epochs. Although the performance of our method is slightly lower than most contrastive learning-based methods (*e.g.*, DINO, MoCo v3 and iBOT), our method only introduces a lightweight frequency decoder without involving complicated contrastive sample selection or data augmentation. Moreover, this mitigation phenomenon on the gap between MIM and contrastive learning also demonstrates the effectiveness of reconstructing the global semantic information.

Transfer Learning Experiments

Object Detection on COCO. To verify Ge^2-AE's transferability, we benchmark it on object detection with COCO (Lin et al. 2014). Following the baseline method MAE (He et al. 2021), the IN1K pre-trained encoder is adopted for initializing the ViT-like backbone of Mask R-CNN (He et al. 2017) framework. Besides, FPN (Lin et al. 2017) is adapted to it (Refer to MAE (He et al. 2021) for more details). We use AdamW (Loshchilov and Hutter 2017) for optimization with initial learning rate $1.6e^{-4}$ and weight decay 0.1.

As vividly shown in Tab. 3, on this visual prediction task, MIM-based methods (BEiT, CAE, MAE and our Ge^2-AE) can present more powerful performance than most contrastive learning-based ones (MoCo v3, PeCo). Although iBOT combining MIM and contrastive learning can achieve the best AP^{box}, our method can achieve the comparable AP^{box} and the best AP^{mask} than other compared methods. Especially to the strong baseline MAE, our method can achieve round 0.6% improvement for either base or large ViT under both protocols, which confirms the necessity of global frequency information in this task.

Method	Pre-train data	AP^{box}		AP^{mask}	
		V-B	V-L	V-B	V-L
Supervised (He et al. 2021)	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3 (He et al. 2021)	IN1K	47.9	49.3	42.7	44.0
BEiT (He et al. 2021)	IN1K+DALLE	49.8	53.3	44.4	47.1
iBOT (Zhou et al. 2021b)	IN1K	**51.2**	-	44.2	-
PeCo (Dong et al. 2021)	IN1K	43.9	-	39.8	-
CAE (Chen et al. 2022)	IN1K	49.2	-	43.3	-
MAE (He et al. 2021)	IN1K	50.3	53.3	44.9	47.2
Ge^2-AE (Ours)	IN1K	51.0	**53.6**	**45.3**	**47.8**

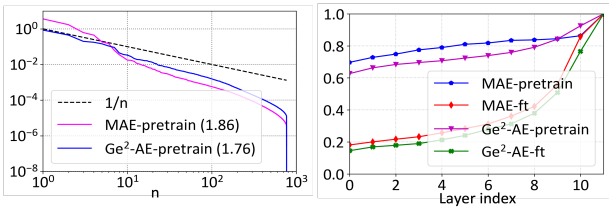
Table 3: Results on COCO object detection and instance segmentation with Mask R-CNN adopting ViT-base (V-B) and -large (V-L) models.

Ablation Study

Configuration	Pixel decoder		Freq. decoder		FT.	LP.
	Pix-re	Freq-con	Freq-re	Pix-con	Top-1	Top-1
Ge^2-AE w/ Pix.	✓	✗	✗	✗	83.6	68.0
Ge^2-AE w/o FD.	✓	✓	✗	✗	83.8	68.3
Ge^2-AE w/ Freq.	✗	✗	✓	✗	82.2	53.9
Ge^2-AE w/o PD.	✗	✗	✓	✓	82.9	55.7
Ge^2-AE w/o con.	✓	✗	✓	✗	84.4	72.5
Ge^2-AE	✓	✓	✓	✓	**84.8**	**75.3**

Table 4: Ablation study of Ge^2-AE on IN1K dataset based on the ViT-B architecture. "PD." and "FD." represent pixel and frequency decoders while "-re" and "-con" denote reconstruction and constraints. "FT." and "LP." are short for fine-tuning and linear probing settings.

To better investigate the effectiveness of different components in our proposed Ge^2-AE, we conduct juxtaposing ablation studies on IN1K dataset under both end-to-end fine-tuning (FT) and linear probing (LP) settings, reported in Tab. 4. The " Ge^2-AE w/ Pix." degrades to the baseline MAE method, where only pixel reconstruction is preserved. By appending frequency constraint ("Ge^2-AE w/o FD."), both fine-tuning and and linear probing Top-1 accuracies witness round 0.3% increase. However, when we remove PD. and only keep FD. preserved but without pixel constraint in Ge^2-AE w/ Freq., there is a 1.4% drop on fine-tuning performance while linear probing suffers more drastically performance devastation. We attribute it to the "over-smoothing" problem, in which the model over-concerns the global se-



(a) Power Law curves (b) CKA Similarity curves

Figure 3: Power Law curves and CKA similarity curves of MAE and our Ge^2-AE after pre-trained and fine-tuned. All experiments are performed on IN1K validation set with ViT base model adopted as encoder.

mantic information while some informative details are abandoned. Although it can be alleviated by appending the pixel constraint (Ge^2-AE w/o PD.), the performance of linear probing setting is still undermined. To further explain this phenomenon, under fine-tuning setting, the contents contained in the lost frequencies (often the high-frequency) can be hunted back according to F-principle (Xu et al. 2019), whereas no chance for linear probing setting. As a full version, Ge^2-AE can fully leverage the information from both pixel and frequency space, which achieves promising results for both FT and LP settings.

Analysis on the Learned Representation

In this subsection, we will further investigate the representation quality learned by Ge^2-AE quantitatively.

Power Law Analysis. According to the recent research in the vision neuroscience (Stringer et al. 2019) and the Machine Learning (Nassar et al. 2020) field, the eigenspectrum of the deep feature covariance matrix often follows a Power Law, of which the coefficient α is strongly correlated to the robustness and generalization of representation. Specifically, we calculate the eigenspectrum of the empirical feature covariance matrix of the pre-trained encoder and estimate its coefficient α as:

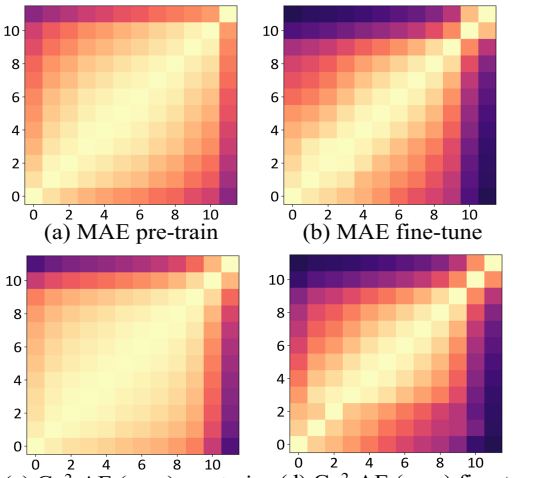
$$\Sigma_N\left(\mathbf{f}_\theta\right)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_\theta\left(x_i\right)\mathbf{f}_\theta\left(x_i\right)^\top, \tag{9}$$

where $\mathbf{f}_\theta(\mathbf{x})$ is the representation. After applying spectral decomposition, the eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$ are obtained, which are all nonnegative and follow Power Law:

$$\lambda_j \propto j^{-\alpha}. \tag{10}$$

Here, α is the slope of the Power Law. As suggested by work (Nassar et al. 2020), *α close to 1 indicates the representation is inclined to exhibit robustness and good generalization.* As demonstrated by Power Law curves in Fig. 3(a), the representation pre-trained by our Ge^2-AE have α value (blue curve, 1.76) closer to 1 (corresponding to dotted line) than that of MAE (magenta curve, 1.86), which confirms the superior quality of the learned representation by our method.

Representation Structure and Evolving Speed. To further study the "*representation structure*" learned in our Ge^2-AE, we plot the Centered Kernel Alignment (CKA) similarities (Kornblith et al. 2019) between all pairs of layers across MAE and our Ge^2-AE after pre-trained or fine-tuned in Fig. 4. We can observe that the fine-tuned features



(a) MAE pre-train (b) MAE fine-tune
(c) Ge^2-AE (ours) pre-train (d) Ge^2-AE (ours) fine-tune

Figure 4: CKA similarities between all pairs of layers across MAE and our Ge^2-AE trained on IN1K. The horizontal axes of heatmaps and vertical axes indexing the layers from input to output. Lighter color indicates higher value. Best viewed in color and zoom in.

(Fig. 4(b) and (d)) often present more compact pattern than pre-trained ones, *i.e.*, merely neighboring several layers can have high similarities when supervision information participating learning. As both MAE and our Ge^2-AE adopting the encoder with same architecture (ViT-B), the representation structures are somewhat similar, but our method can still exhibit more compact structures (Fig. 4(c) and (d)) than MAE (Fig. 4(a) and (b)) for both training settings.

To profile the "*evolving speed*" of learned representations, we also present the CKA similarity curves in Fig. 3(b), where similarities are computed between each layer and the last one. As concluded in previous work (Zhou et al. 2021a; Xu et al. 2021), if the features change slowly when traversing the model layers, the model could perform inferior to the ones with faster feature evolving speed. By comparison, both pre-trained and fine-tuned features of Ge^2-AE witness faster evolving speed than MAE, which further verifies the effectiveness of our method.

Learning Patterns of Ge^2-AE

Visualization of FSP Weights. To explore the learning pattern of Frequency Decoder (FD) of our Ge^2-AE, we in Fig. 6 visualize the first 24-channel weights Ω in Fourier Spectrum Perceiver (FSP) of each FD block (Blk 1 $\sim$ Blk 8). In the first several FD blocks nearing encoder output, FSP weights exhibit diversified patterns, *i.e.,* different channels of FSP emphasize different frequency components where each one corresponds to an exclusive global informative content. Along with block number increasing, the patterns of FSP weights become reasonably sparse, because only those useful frequencies are preserved for reconstruction task during this procedure of global high semantic extraction. Thanks to the rich patterns serving as candidates, more effective information hidden in the frequency can be leveraged. Conclusively, we believe this "from-dense-to-

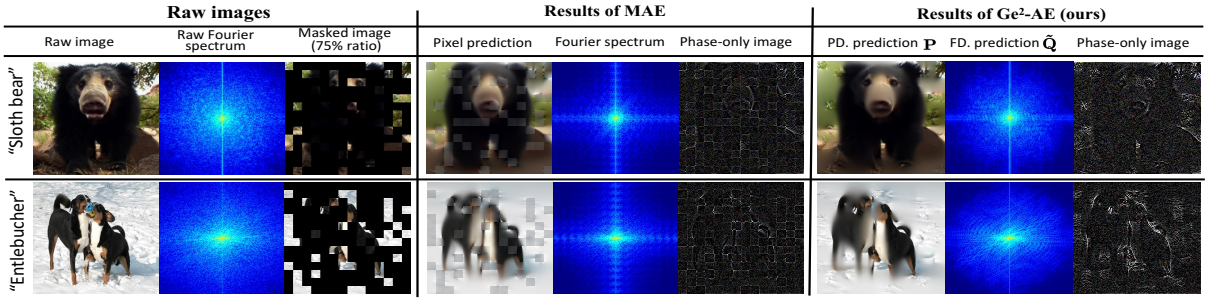


Figure 5: Visualizations of the predicted results from MAE and our Geminated Gestalt Autoencoder (Ge^2-AE) pre-trained on IN1K dataset. "PD." and "FD." are short for pixel decoder and frequency decoder respectively. Their outputs are $\mathbf{P}$ and $\tilde{\mathbf{Q}}$, corresponding to the counterparts in Fig. 2(a). Our method can yield results with more necessary global frequency and local details than MAE to overcome the "over-smoothing" issue. Best viewed in color and zoom in.

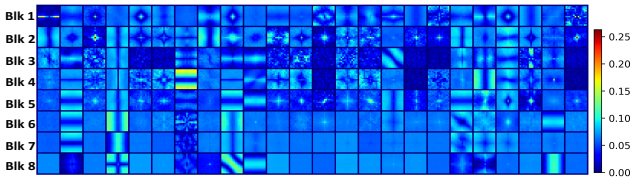


Figure 6: Weight visualization of Fourier Spectrum Perceiver (FSP) in frequency decoder of Ge^2-AE. Best viewed in color and zoom in.

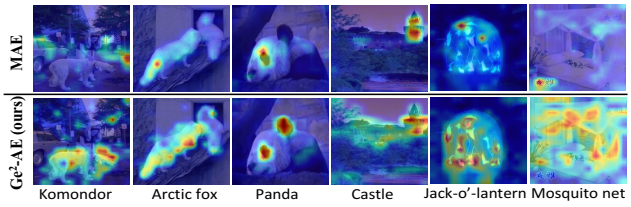


Figure 7: Class activation attention maps on classification task under linear probing setting. Best viewed in color.

sparse" pattern evolution is one of the keys to solve "over-smoothing" problem.

Qualitative Analysis on Predictions. In Fig. 5, we visualize the predictions of baseline MAE (He et al. 2021) and our Ge^2-AE adopting ViT-B as encoder with images masked in 75% ratio. More concretely, the results in "Fourier spectrum" column from MAE are obtained by applying 2D-FFT on recovered images in "pixel prediction" column, while "phase-only images" are obtained by setting the amplitude of Fourier spectrum to constant and restored to pixel image through 2D-IFFT.

Visually, the "phase-only images" of MAE present corrupted object contour, and the corresponding "Fourier spectrum" may contain unwanted frequency components but with deserved local details ignored. We attribute it to the inherent mechanism in MAE, where only the pixels at masked locations participate the pixel loss computation, while those unmasked ones are also effected uncontrollably. Although MAE attempts to alleviate it by calculating loss for all pixels, the model is inclined to directly copy the original pixel values of unmasked ones while masked pixels are less focused on, resulting in the performance decrease.

Comparatively, the predictions of pixel decoder ("PD. prediction $\mathbf{P}$") in our method (Fig. 5) can generate pixel image without unmasked local details lost, *e.g.,* the eyes of "Sloth bear". Correspondingly, the predictions of frequency decoder can output Fourier spectrum maps $\tilde{\mathbf{Q}}$ with more global frequencies preserved. Consequently, the phase-only images exhibit more reasonable semantic properties with more obvious and holistic object contours.

Class Activation Maps of Linear Probing. Based on interpretability tool (Chefer, Gur, and Wolf 2021), in Fig. 7, we visualize the attention maps of class activation from the model trained under linear probing setting. Obviously, the attentions of MAE often concentrate on the more local regions, which may caused by the missing of proper global semantic information. In a stark contrast, our method can present more reasonable attention regions with more scalability. For example, MAE only focuses on the single eye region of "panda" while ours highlights both eye and ear regions with black color. Besides, for the "mosquito net" case, MAE wrongly pays attention to the logo while our method correctly attend to the whole object, which is attributable to the superior global semantic abstraction ability.

Conclusion and Future Work

We in this paper rethink the Masked Image Modeling (MIM) for visual pre-training task from frequency perspective and propose a novel Ge^2-AE with dual decoders to reconstruct image contents from both pixel and frequency spaces. Extensive experiments on downstream visual tasks verify the superior performance of it than state-of-the-arts. A series of analytic experiments are also conducted to further investigate and explain our model from both quantitative and qualitative aspects. We believe shifting perspective to frequency domain could be an enlightening attempt to the community. However, there may still exist redundancy in frequency. We suppose the frequency information could be encoded more compactly, which can be achieved by introducing more deductive method, such as quantization, in the future.

References

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 782–791.

Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2022. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; and Yu, N. 2021. PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. *arXiv preprint arXiv:2111.12710*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, Y.; Dong, L.; Bao, H.; Wang, X.; and Wei, F. 2022. Corrupted Image Modeling for Self-Supervised Visual Pre-Training. *arXiv preprint arXiv:2202.03382*.

Hansen, B. C.; and Hess, R. F. 2007. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24(7): 1873–1885.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13919–13929.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 3519–3529. PMLR.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, H.; Jiang, X.; Li, X.; Bao, Z.; Jiang, D.; and Ren, B. 2021. NomMer: Nominate Synergistic Context in Vision Transformer for Visual Recognition. *arXiv preprint arXiv:2111.12994*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nassar, J.; Sokol, P.; Chung, S.; Harris, K. D.; and Park, I. M. 2020. On 1/n neural representation and robustness. *Advances in Neural Information Processing Systems*, 33: 6211–6222.

Nussbaumer, H. J. 1981. The fast Fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, 80–111. Springer.

Oppenheim, A.; Lim, J.; Kopec, G.; and Pohlig, S. 1979. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, 632–637. IEEE.

Oppenheim, A. V.; and Lim, J. S. 1981. The importance of phase in signals. *Proceedings of the IEEE*, 69(5): 529–541.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Piotrowski, L. N.; and Campbell, F. W. 1982. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3): 337–346.

Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.; Bengio, Y.; and Courville, A. 2019. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 5301–5310. PMLR.

Stringer, C.; Pachitariu, M.; Steinmetz, N.; Carandini, M.; and Harris, K. D. 2019. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765): 361–365.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8684–8694.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and Feichtenhofer, C. 2021. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *arXiv preprint arXiv:2112.09133*.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2021. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*.

Xu, Y.; Zhang, Z.; Zhang, M.; Sheng, K.; Li, K.; Dong, W.; Zhang, L.; Xu, C.; and Sun, X. 2021. Evo-vit: Slow-fast token evolution for dynamic vision transformer. *arXiv preprint arXiv:2108.01390*.

Xu, Z.-Q. J.; Zhang, Y.; and Luo, T. 2022. Overview frequency principle/spectral bias in deep learning. *arXiv preprint arXiv:2201.07395*.

Xu, Z.-Q. J.; Zhang, Y.; Luo, T.; Xiao, Y.; and Ma, Z. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*.

You, Y.; Gitman, I.; and Ginsburg, B. 2017. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*.

Zhou, D.; Shi, Y.; Kang, B.; Yu, W.; Jiang, Z.; Li, Y.; Jin, X.; Hou, Q.; and Feng, J. 2021a. Refiner: Refining self-attention for vision transformers. *arXiv preprint arXiv:2106.03714*.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021b. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.